Unsupervised Machine Translation in Real-World Scenarios

AUTHORS

Ona de Gibert, Iakes Goenaga, Jordi Armengol-Estapé, Olatz Perez-de-Viñaspre, Carla Parra, Marina Sánchez-Torrón, Marcis Pinnis, Gorka Labaka, Maite Melero

ona.degibert@bsc.es

AFFILIATIONS

Barcelona Supercomputing Center, Barcelona, Spain
HiTZ zentroa - Ixa, UPV/EHU, Donostia, Spain
RWS Language Weaver, Dublin, Ireland
Unbabel, Lisbon, Portugal
Tilde, Riga, Latvia

ABSTRACT

In this work, we present the work that has been carried on in the MT4All CEF project and the resources that it has generated by leveraging recent research carried out in the field of unsupervised learning. In the course of the project, 18 monolingual corpora for specific domains and languages have been collected, and 12 bilingual dictionaries and translation models have been generated. As part of the research, the unsupervised MT methodology based only on monolingual corpora [3] has been tested on a variety of languages and domains. Results show that in specialised domains, when there is enough monolingual in-domain data, unsupervised results are comparable to those of general domain supervised translation, and that, at any rate, unsupervised techniques can be used to boost results whenever very little data is available.

WANT TO KNOW MORE?

READ OUR PAPER!

IN **LOW-RESOURCE SCENARIOS**, COMPLETELY **UNSUPERVISED MACHINE TRANSLATION** TENDS TO YIELD POOR RESULTS, EXCEPT WHEN THE AMOUNT OF **IN-DOMAIN** MONOLINGUAL DATA IS BIG ENOUGH.

MOTIVATION

- Parallel data is not always available, specially for low-resource languages.
- Few open-source systems for low-resource scenarios, using supervised techniques. [6]
- Unsupervised Machine Translation (MT) has achieved impressive results. [2, 3, 4]
- Unsupervised methods [2] use standard benchmarks for comparison, with high-resource language pairs.
- Expansion of unsupervised techniques [2] to further domains and languages.

PUBLISHED RESOURCES

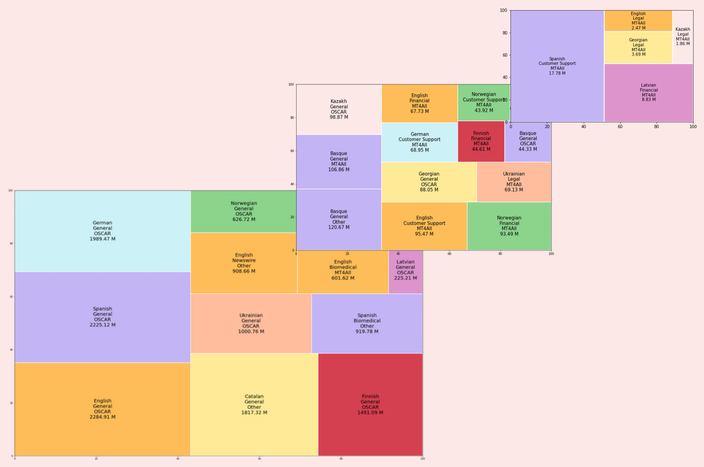
Our contributions:

- Release of 18 monolingual corpora
- Release of bilingual dictionaries, word embeddings and translation models.

BIOMEDICAL EN<>ES	CUSTOMER SUPPORT EN<>DE, ES, NO	FINANCIAL EN<>FI, LT, NO
GENERAL EN<>CA	LEGAL EN<>KA, KK, UK	NEWSWIRE EN<>EU

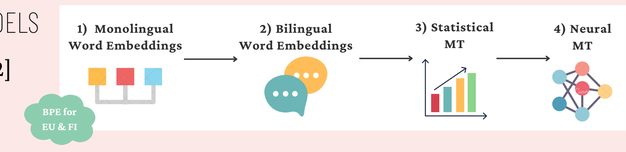
METHOD AND RESOURCES

TRAINING CORPORA



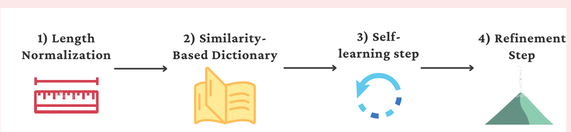
TRANSLATIONS MODELS

We use Monoses [2]



WORD EMBEDDINGS

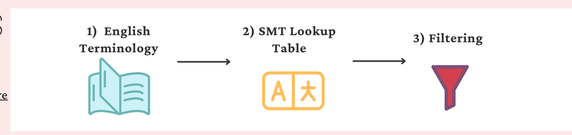
We use Vecmap [1]



BILINGUAL DICTIONARIES

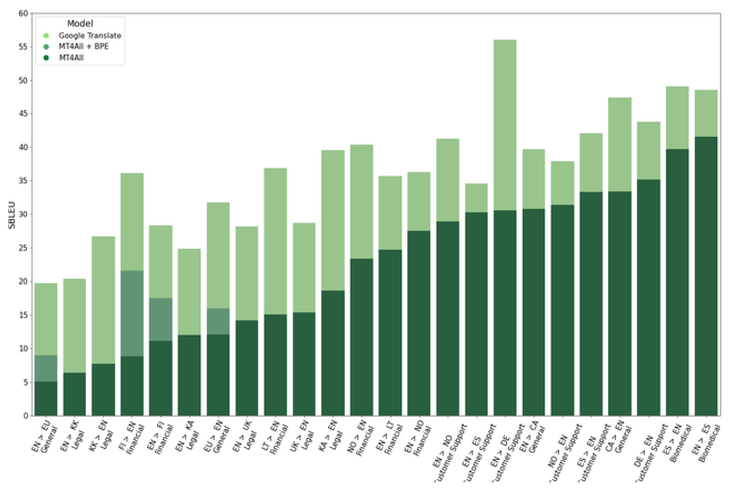
We use TermSuite

https://github.com/termsuite/termsuite-core



RESULTS

AUTOMATIC EVALUATION

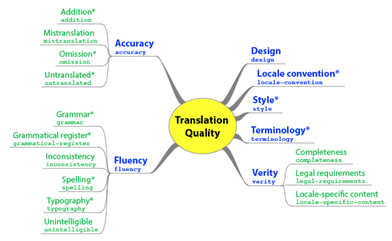


- The unsupervised models do not achieve parity with the supervised model.
- Factors to take into account:
  - the size of in-domain corpora
  - the size of general corpora
  - the distance from English
  - the level of complexity
- The BPE segmentation boosts the results.

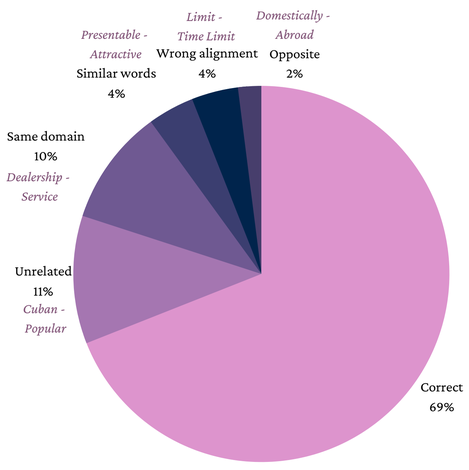
HUMAN EVALUATION

Three domains:

- **Customer Support (ALL)**
  - Named Entities: *Barney > Rosendo*
  - Localization: YYYY-MM-DD
  - Formal/ Informal Registers
  - Duplications: *Takk, Takk*
  - Parsing of URLs: *www.acme.no / blogg*
- **Financial (EN>LT)**
  - 100 sentences
  - Mistranslations
  - Named Entities
  - Omissions
  - Typography
  - MQM Score*: 350
- **Biomedical (EN>ES)**
  - Full test set
  - Mistranslation
  - Repetitions
  - Named Entities
  - No fluency errors
  - MQM Score*: 138



MULTIDIMENSIONAL QUALITY METRICS FRAMEWORK



INSPECTION OF 100 RANDOM ENTRIES OF THE EN>LT DICTIONARY

*Reference MQM Score for humans: 50

CONCLUSIONS

- In low-resource scenarios, completely unsupervised systems tend to yield poor results, except when the amount of in-domain monolingual data is big enough to compensate.
- Human evaluation is necessary, better than automatic metrics, to capture the output of an MT system.
- In real-world scenarios, there is always some access to parallel data or it can be created synthetically by triangulation or other methods.

REFERENCES

1. Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 789–798, 2018.
2. Mikel Artetxe, Gorka Labaka, and Eneko Agirre. An effective approach to unsupervised machine translation, 2019.
3. Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041, 2017.
4. Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043, 2017.
5. Matt Post. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.
6. Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for theWorld. In Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.